(1)

Research Note 88-42

STRUCTURING KNOWLEDGE RETRIEVAL: AN ANALYSIS OF DECOMPOSED QUANTITATIVE JUDGEMENTS

Donald MacGregor, Sarah Lichtenstein,
and Paul Slovic
Perceptronics, Inc.

for

Contracting Officer's Representative
Michael Drillings

ARI Scientific Coordination Office. London
Milton D. Katz, Chief

BASIC RESEARCH OFFICE
Michael Kaplan, Director

U. S. Army
Research Institute for the Behavioral and Social Sciences

June 1988

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

Research accomplished under contract
for the Department of the Army

Perceptronics, Inc.

Technical review by

Dan Ragland

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ARI Research Note 88-42

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Structuring Knowledge Retrieval: An Analysis of Decomposed Qualitative Judgements

5. TYPE OF REPORT & PERIOD COVERED
Interim Report
April 86 - April 87

6. PERFORMING ORG. REPORT NUMBER
- -

7. AUTHOR(s)
Donald MacGregor, Sarah Lichtenstein, and Paul Slovic

8. CONTRACT OR GRANT NUMBER(s)
MDA903-83-C-0198

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91367

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q161102B74F

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600

12. REPORT DATE
June 1988

13. NUMBER OF PAGES
47

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
- -

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
- -

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
- -

18. SUPPLEMENTARY NOTES

Michael Drillings, contracting officer's representative

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Behavioral Decision Analysis
Knowledge Elicitation
Knowledge Acquisition
Value Estimation
Decision Support

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Subjects were asked to estimate the answers to sixteen questions concerning uncertain quantities like "How many people are employed by hospitals in the U.S.?" under five different aiding conditions. The most-aided group (Full Algorithm) was given a complete algorithm and asked to make estimates for all the parts of the algorithm and to combine the parts as indicated to arrive at an estimate of the desired quantity. The second group (Partial Algorithm) was given the same algorithm without indications of how to
(OVER)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Structuring Knowledge Retrieval:

An Analysis of Decomposed Quantitative Judgments

Donald MacGregor, Sarah Lichtenstein, and Paul Slovic

Abstract

Subjects were asked to estimate the answers to sixteen questions concerning uncertain quantities like "How many people are employed by hospitals in the U.S.?" under five different aiding conditions. The most-aided group (Full Algorithm) was given a complete algorithm and asked to make estimates for all the parts of the algorithm and to combine the parts as indicated to arrive at an estimate of the desired quantity. The second group (Partial Algorithm) was given the same algorithm without indications of how to combine the parts. After making estimates of the parts, these subjects then estimated the desired quantity. The third group (List & Estimate) were asked to list components or factors they thought were relevant, make an estimate of each item on their list, and then estimate the desired quantity. The fourth group (List) were asked to make such a list, but they were not asked to make estimates of each item before making an estimate of the desired quantity. The fifth group received no aid. The results generally showed improved performance in terms of both accuracy and consistency across subjects with increasing structure of the aid. Generalization of these results to practical estimation situations is possible but limited by the need, in real situations, for the estimator to develop the algorithm, a task that was here done by the experimenters.

Structuring Knowledge Retrieval

Decision makers often need to know the value of a particular quantity, such as "how much money did our family spend on food last year?" or "what is the Soviet troop strength in Cuba?" Sometimes the quantitative value is readily obtainable from a computer database, encyclopedia, reference source or expert. In other cases, one may have read or been told the value so that it can be accurately and confidently retrieved from memory when it is needed.

There are some quantities, however, that may not exist in any reference sources, or may be available only in sources that are difficult to locate, prohibitively costly to obtain, take too long to use, or contain only vague and partial knowledge. Faced with this situation, the best one can do is make an estimate of the needed quantity based on the information resources at one's disposal. In principle, those resources could extend to decision support systems, computer databases, information libraries, and the judgments of experts. In practice, however, estimates are often based on whatever relevant knowledge an estimator is able to obtain.

Even when a computer database is available, its contents may contain information related to the quantity called for, but not the quantity itself. Indeed, making full use of information systems entails understanding both the capacity of the system to provide direct answers to questions and its capacity to provide relevant information for questions it cannot answer directly. Effective

querying of the system in the latter case requires a careful structuring of the user's information requirements, the absence of which can lead to inefficient use of computer resources, incomplete information retrieval, or erroneous results. How well one is able to exploit computerized information resources in such situations is in part dependent upon one's ability to structure whatever bits of information the system can provide in a form that is meaningful for the task of estimating the quantity in question. While the substantive contents of that structuring are potentially available from an information system, the form of the structuring is usually left to the idiosyncracies of the individual user. Studying the behavioral properties of various methods of structuring information helps shed light on the relative performance that might be expected from individuals when they are provided specific guidance on how to approach organizing quantitative estimation problems.

Unfortunately, we have as yet no general theory of how knowledge structuring approaches should be developed. We do, however, have some suggestions that point the way to principles that could prove useful in aiding quantitative judgments. Raiffa (1968), for example, advises to ". . . decompose a complex problem into simpler problems, get one's thinking straight in these simpler problems, paste these analyses together with a logical glue, and come out with a program for action for the complex problem" (pg. 271).

On the other hand, Hammond and his colleagues have argued that

analytic judgment strategies exhibit different cognitive properties than do intuitive ones (e.g., Hammond, 1981; Hammond, Hamm, Grassia, & Pearson, 1983). Their general theory of a cognitive continuum predicts that specific characteristics of a task, including the characteristics and organization of information, can induce different modes of cognition and performance. Although decomposition is intended to improve the tractability of a task, cognitive continuum theory predicts that the induction of an analytic mode of thinking may produce systematic biases in the types of judgmental errors that are made. For example, Peters, Hammond and Summers (1974) found that intuitive judgments are, for the most part, approximately correct, whereas analytically based judgments exhibit a large number of precisely correct responses with a few extremely large errors.

The exact tradeoff to make between analytical sophistication and intuitive simplicity in a knowledge-structuring approach for aiding the estimation of uncertainty quantities is difficult on purely theoretical grounds. This paper takes the position that direct tests of plausible structuring aids can provide valuable insights into their behavioral properties.

Approaches to structuring knowledge retrieval for the sake of estimating an uncertain quantity can vary in form and elaboration. The simplest approach is to consider what one knows about a quantity of interest and intuitively divine an estimate that seems reasonable in light of whatever knowledge comes to mind. This

wholistic approach to estimation relies heavily on the power of unaided human cognition. To its advantage, it is inexpensive, portable, and represents the way in which people routinely deal with estimation problems, thus achieving a high degree of psychological compatibility. To its disadvantage, memory could prove to be too impoverished a resource on which to base an estimate and may provide no indication of which information might prove useful. Even when seemingly useful knowledge is retrieved, there may be no way of knowing how to combine disparate pieces of knowledge into a global estimate. Furthermore, research on the psychology of human judgment has repeatedly demonstrated that simplifying cognitive strategies can lead to systematic judgmental biases. Judgments that use convenient starting points as a basis for estimating the magnitude of a quantity can exhibit a tendency for insufficient movement away from an initial value (Tversky & Kahneman, 1974). The ease with which instances of a phenomenon are retrieved from memory may be influenced by recency, salience and vividness, factors that have no bearing on the rate with which the phenomenon occurs in the real world. Thus, when the availability of relevant events in memory is used as a basis for estimating an uncertain quantity, those estimates may be biased by cognitive processes that are a natural part of knowledge retrieval (Tversky & Kahneman, 1973).

An alternative to the wholistic, intuitive approach is analysis or decomposition. This involves breaking up or

decomposing a problem into a series of sub-problems or components, each of which can be understood more easily and operated on separately. The components are then assembled according to a prescribed set of combination rules to yield a solution, estimate or prediction. Decomposition is a divide-and-conquer approach that assumes the components of a problem to be more understandable and tractable than the undecomposed problem.

Analysis and decomposition have been employed in a wide variety of problem areas. For example, decision analysis, a methodology for choosing in situations involving uncertainty (Raiffa, 1968), partitions a decision problem into actions and outcomes. Each outcome has an associated payoff amount and probability which are analyzed to determine the optimal course of action. The decomposition principle has also been applied to human judgment. Typically, a judgment task is decomposed into a number of relevant cues. Cues are generally combined via an additive linear model (e.g., Einhorn, 1972), with the importance of each cue reflected by an associated numerical weight, usually derived through multiple regression.

A variant of the decompositional approach is the algorithm. An algorithm is a series of steps or operations that, when sequentially applied, produce a solution to a problem. Though the approach is popularly associated with the computer science field (e.g., Goodman & Hedetniemi, 1977), it has formal application in such diverse areas as teaching rules of inference and deduction

(Landa, 1976), communication of procedural knowledge (Horabin & Lewis, 1978; Wright & Reid, 1973) and judging human performance (Lyness & Cornelius, 1982).

Essentially, algorithms work by providing an unambiguous procedure for solving problems. They help structure what is known about a problem, point out what is not known, and specify the rules by which information should be combined. Since the combination of information is mechanical, algorithms have the potential for high reliability; different individuals using the same algorithm should arrive at very nearly the same solution.

Singer (1971) illustrated the use of algorithmic decomposition to estimate the amount of money per year taken in muggings, robberies, and burglaries by heroin addicts in New York City. Using an algorithm with components he could estimate more accurately, such as the population of the city, the number of all reported burglaries, and the number of addicts in prison, he arrived at an estimate of $250 million stolen per year, far smaller than estimates previously suggested.

Although Singer's article is a compelling (and clever) argument for the advantages of algorithmic decomposition, experimental evidence is needed before the technique can be prescribed as an estimation aid. Part of that evidence should involve comparisons of algorthmic decomposition with other, more simplified, structuring approaches and with direct, global estimation. As the first stage of such a program, Armstrong,

Dennison, and Gordon (1975) showed that, for five uncertain quantities, estimates given by subjects to the sub-parts of an algorithm, subsequently combined by the experimenters, yielded greated accuracy than did direct, global estimates of the target quantities.

What these research results suggest is that increasing levels of problem structuring tend to improve the quality of numerical estimates, and that full, complete algorithms, with users carrying out both estimation and combination of components should outperform both direct estimation and estimation aided by partial algorithms. One can also hypothesize that estimation aids calling for users to structure their own problems should do better than direct, global estimation, but not as well as algorithm-aided estimation.

An important question to ask of any judgment aid is what effect it has on users' confidence in the accuracy of estimates. A robust finding from research on the psychology of confidence is that people have an exaggerated belief in how much they know. For example, Lichtenstein and Fischhoff (1977) found that when people are asked to assess the probability that an answer they have chosen to a general knowledge question is the correct answer, the proportion of the time they are right for all items assigned the same probability value is typically too low; their probability assessments exhibit overconfidence.

Appropriateness of confidence can be precisely determined for probability assessments but it is less specifically gauged for

estimates of uncertain quantities. How confident one is in the accuracy of a point estimate of a continuous quantity depends in part on the degree of precision one is interested in. Thus, for a relatively broad interval around a point estimate one should express more confidence that the correct answer would fall within the interval than within a narrower interval. For the sake of comparing confidence in the accuracy of estimates produced by a number of structuring approaches, however, a comparison on a relative scale of confidence may have to suffice. Structuring techniques that elicit greater relative confidence should also elicit greater relative accuracy for that confidence to be at all appropriate.

Method

Overview of the Study

The present study involved five groups of subjects performing an estimation task. Four of the groups were aided with one of four levels of knowledge structuring. The fifth group performed the same task but received no estimation aiding. The quantities estimated were posed as almanac questions of the type "How many cigarettes are consumed in the U. S. in a year?"

Full Algorithm. At this, the highest level of structuring, each estimation problem was decomposed into a complete algorithm. The subjects made estimates of each of the component parts and

combined their estimates according to specified arithmetic operations. For example,

<u>How many cigarettes are consumed in the U. S. in a year?</u>

A. What is the population of the U. S.?

B. About what proportion of the population smokes?

C. Multiply (A) x (B) to get number of smokers.

D. How many cigarettes does the average smoker consume per day?

E. Multiply (C) x (D) to get number of cigarettes consumed in the U. S. in a day.

F. How many days are there in a year?

ANSWER: Multiply (E) x (F)

<u>Partial Algorithm</u>. At this level of structuring, estimates were made for each of the components of the full algorithm, followed by an estimate of the target quantity. Unlike full algorithms, however, no rules by which to combine the component estimates were given. This approach provided a less complete structuring of the estimation problems than did the Full Algorithm condition.

<u>List & Estimate</u>. In this condition, subjects provided their own problem structuring. Before estimating a target quantity, they first listed components or factors that they believed were relevant to estimating the target quantity; they then estimated each of the components they had listed. Unlike the Full Algorithm and Partial Algorithm conditions, this condition did not limit respondents to a

particular problem representation, but still called for component estimates.

List. Subjects were instructed only to list components or factors they believed were relevant to estimating the target quantity. They were not asked to make estimates of any items they listed, nor were they asked to combine information in a particular manner, though they were not restricted from doing so. An estimate was then made of the target quantity.

Unaided. No structuring was provided to subjects in this condition. Each of the target quantities used in the study was estimated directly as a control against which to compare the performance of the four structuring approaches.

Estimation Questions

Sixteen almanac-type questions were used for each of the five conditions. The correct answers to the question varied in magnitude from 350 (How many practicing physicians are there in Lane County?) to 604.1 billion (How many cigarettes are consumed in the U. S. in a year?) Table 1 lists all the estimation problems and the correct answer[1] for each problem.

Insert Table 1 About Here

The majority of questions were based on quantities contained in statistical almanacs. Other questions were based on information obtained from local sources and related to topics such as water

use, welfare payments, and the like. In all cases, the estimation problems related to quantities which most, if not all, subjects were unlikely to have much direct experience with or to have thought about extensively. All of the problems, however, related to topics about which people were likely to have some relevant knowledge.

The full set of 16 algorithms is shown in Table 2. For the sake of brevity, only the algorithm steps requiring subjects to make component estimates are provided. Intermediate arithmetic steps are omitted.

Insert Table 2 About Here

The set of problems was chosen arbitrarily from a number of resources and was selected to provide subjects with estimation questions having a wide range of true answers. No attempt was made to create systematically a problem set that represented a population of problem types in terms of dimensions such as number of steps in the algorithm, computational complexity or ease of estimation of component estimates.

<u>Procedure</u>

The task was introduced as follows:

> We are interested in how accurately people can estimate unfamiliar quantities. On the following pages you will be asked to estimate answers to a number of

> quantities taken from almanacs and other sources of statistics. You probably won't know the true values for most or all of these quantities.

The subjects in the Full Algorithm condition then received an example problem illustrating the method, followed by a set of estimation problems to complete. Partial Algorithm subjects were given a brief introduction to the task and told that before estimating each quantity, they would make some estimates of other related quantities, and that the accuracy of their estimate might be improved by giving careful consideration to the related quantities.

In the List & Estimate condition, subjects were instructed first to list some quantities they believed one should consider in making an estimate. These could be factors or components that would be useful in arriving at an estimate. They were then asked to give their best estimate for each quantity they listed. After listing and estimating, they gave their best estimate of the target quantity, reconsidering the component quantities they listed. The List condition asked subjects only to list some things that they thought one should consider in making an estimate. Again, this could be a list of factors or components that could be useful in arriving at an estimate. They were not asked to provide estimates for anything they listed. Unaided subjects were simply given a general introduction to the task and a list of quantities to estimate.

After completing each estimation problem, subjects indicated their confidence in the accuracy of their estimate on a seven-point scale ranging from "very unconfident" (= 1) to "very confident" (= 7). The same confidence ratings were made in all five conditions.

The task was presented to the subjects untimed, allowing them to work on each estimation problem as diligently as they wished.

Subjects

A total of 514 subjects recruited through an ad in the University of Oregon student newspaper participated in the study. The subjects also performed several other, unrelated judgment and decision-making tasks. They were each paid $5 for their participation. Subjects in the Unaided condition (N=45) received the entire set of 16 estimation problems. The set of 16 estimation problems presented to the remaining groups was divided into smaller sets and administered to separate subject groups. For the List and Partial Algorithm conditions the problem set was divided into two subsets of eight problems each and administered to separate groups of subjects (N = 45 and 41). The problem set was divided into three subsets (6, 5, 5) for the Full Algorithm condition and the same procedure followed (N = 36, 38, and 45). Four subsets of four problems each were created for the List & Estimate condition (N = 43, 44, 45, 46).

Results

Accuracy of Estimation

The geometric means of estimates produced under each of the structuring conditions along with the correct answer for each of the estimation questions are shown in Table 3. They are listed in order of magnitude of the correct answer. The geometric mean was chosen as a summary statistic because it reduces the influence of extreme data points. The estimation task did not imply any bounds around the size of estimates subjects might produce. That property of the task pointed to the geometric mean as the appropriate summary statistic.

Insert Table 3 About Here

Table 3 shows a high degree of variability of the estimates for each problem across the estimation conditions. The ratio of the highest to the lowest geometric mean estimate produced for a particular estimation problem across conditions varied from 1.8 for the "hospital employees" problem to 396.8 for the "cigarettes" problem. On average, the highest and lowest mean estimates for each of the problems varied by a factor of 47 across the five estimation conditions.

The data of Table 3 are summarized in Table 4 as error ratios, that is, the ratio of the geometric mean estimate to the correct answer, or vice versa, such that the result is equal to or greater

than one. These error ratios do not indicate whether subjects had a tendency to over- or underestimate the correct answer. That information is shown in Table 4 by a "+" or "-" for over- and underestimation respectively. Underestimation was the most common directional bias; for only 12 of the 80 data points in Table 4 did overestimation occur, most often for problems having relatively small correct answers.

Insert Table 4 About Here

The median and mean error ratios appear at the bottom of Table 4. These summaries show that the List condition led to better performance than the Unaided condition. However, the List & Estimate condition was not superior to the Unaided condition; indeed, its error ratio was larger for 9 of the 16 questions. A possible explanation for this result lies in the more demanding nature of the List & Estimate task, where the additional effort of providing estimates of quantities subjects listed could have had the effect of focusing their attention away from the global task and onto the details of estimating subquantities, thereby eroding performance. Alternatively, subjects may have been faced with a list of quantities that they did not know how to integrate into a global estimate. The task of intuitive integration may have led to confusion or to erroneous use of what knowledge subjects were able to produce in their lists of information.

Both the Algorithm groups showed distinctly better performance than the other three groups, by avoiding the very large underestimation errors made for the eight largest target quantities. Since both of the Algorithm conditions gave subjects detailed problem representations in organized formats, it may not seem surprising that these two conditions elicited the best performance. However, those representations could have confused subjects had they been considerably different from the way they might have naturally thought about those same problems.

There was no significant difference in performance between the Partial Algorithm group and the Full Algorithm group, although the latter was noteworthy in never erring by a factor greater than 10.

<u>Consistency of Estimation</u>

What one ultimately desires out of a method for aiding estimation is, of course, an improvement in accuracy. A second and, in some circumstances, equally desirable goal is the attainment of consistency. Consistency is desirable on several grounds. First, when a correct answer against which to compare a judgmental estimate is not known, then the quality of that estimate must be assessed by recourse to the properties of the method by which it was produced. If a method has known biases, then those biases can be corrected for, provided the method is known to yield consistent results. Moreover, for a method to be of general use, the quality of its application should depend as little as possible on the idiosyncracies of individual users. It should, for example,

be understood equally well by all who apply it and should not introduce ambiguities into the process it intends to aid.

Consistency is manifest in different forms. One form is the tendency for a method to produce the same results when applied by the same user under identical circumstances. A second form of consistency is reflected in the tendency for a method to produce similar results in the hands of different users. The latter form is more appropriate for the types of knowledge structuring methodologies studied here.

Table 5 summarizes the consistency across subjects for each of the knowledge structuring conditions, as indicated by measures of the range and interquartile range. The range measure used here is the log of the ratio of largest response to the smallest response.[2] Thus, a range of 2.00 means that the largest response was 100 times as great as the smallest response; a range of 6 shows a ratio of largest to smallest of one million. The interquartile range is the log of the ratio of the third quartile response to the first quartile response.

Insert Table 5 About Here

The largest range in Table 5 is 19.00, the range for the Unaided group on the Cigarette question. One subject in that group gave an answer of 100,000. This subject tended to give low answers; three of her responses were within a factor of 10 of the

correct answer and the remaining 13 responses were more than 10 times too small. The largest response to the Cigarette question was:

"Far to (sic) many—1,000,000,000,000,000,000,000,000"

The subject giving this response also gave the largest response in her group to three other questions, University Employees, Baseball, and Alcohol Dollars. Overall, 5 of her responses were within a factor of 10 of the correct answer, 8 were more than 10 times too high, and 3 were more than 10 times too low.

Contrasting these two subjects' responses (10^5 vs. 10^{24}) to the same estimation question presented in the same manner illustrates the range of interpretations individuals can give to an estimation task when the magnitude of the quantity called for is very large. When numbers in the millions or billions are called for and no structuring aid is provided, it may be more natural for people to imagine the answers ideosyncratically and in an impressionistic way ("exceedingly large") than to think of them in strict numerical terms.

Figure 1 shows the distribution of responses for the Unaided group on the Cigarette question and, by way of contrast, the distribution of the Full Algorithm group on the same question. The latter is more typical of the shape of 80 response distributions summarized in Table 5. Note that the Unaided group tended to grossly underestimate the correct answer; only 4 of 44 subjects gave responses greater than 10^{11} (100 billion). This is also

indicated by their error ratio of 393.29 for this question (shown in Table 4). The Full Algorithm group also showed underestimation, but to a lesser extent.

Insert Figure 1 About Here

The bottom row of Table 5 reports the median values of both range statistics for each estimation condition. The range of estimation was greatest for the Unaided condition (median = 7.94) and smallest for Full Algorithm (median = 5.59). That same result is reflected as well in the interquartile range, where Unaided had the largest interquartile range (median = 1.75) and Full Algorithm the smallest (median = 1.30). The interquartile ranges reveal a tendency for less variability (greater consistency) with increasing problem structuring, particularly when subjects were provided with a specific problem representation, as they were in the two Algorithm conditions. The same general pattern appears as well in the overall range values, though not quite as distinctly.

What these data do not reflect is a tendency for greater knowledge structuring and use of problem-specific representations to produce a large clustering of highly accurate estimates (as indicated by the interquartile range) with a small number of extreme outliers (as indicated by the range), relative to more intuitive, global approaches (see Peters et al., 1974). Had that been the case, the magnitude of the interquartile range relative to

the overall range would have been smaller for the Partial Algorithm and Full Algorithm conditions than for the remaining three approaches, none of which provided subjects with analytically oriented problem formats. Instead, the results in Table 5 suggest that as estimation aiding techiques increase the degree to which they structure problem organization and specify information requirements, more consistent estimates across individual users will be produced. Moreover, the increased consistency produced by the aids was not purchased at the cost of reduced accuracy. Across the 16 questions, the interquartile range contained the correct answer 7, 8, 8, 11, and 12 times (out of 16) for the Unaided, List, List & Estimate, Partial Algorithm, and Full Algorithm conditions, respectively.

Recomputed Responses

The data reported so far were based on the actual estimates of the target quantities as provided by subjects. Subjects in the Full Algorithm condition often made arithmetic errors in arriving at their target estimates. We produced new estimates of the target quantities by correctly combining each subject's estimates of the component quantities. Subjects in the Partial Algorithm condition were never told the appropriate arithmetic steps for combining the components. For these subjects new estimates were produced by combining their estimates of the components via the algorithm.

We also computed estimates based on the principle of bootstrapping (see Dawes & Corrigan, 1974). To bootstrap the

target estimates, each component in the algorithm was assigned a numerical value equal to the median of the estimates made by the subjects for that component. These median component estimates were combined via the algorithm. The error ratios associated with these two new estimates, one based on the experimenters' computations and the other on bootstrapping, are compared with the original error ratios for the Partial and Full Algorithm groups in Table 6.

Insert Table 6 About Here

For both groups, these revised estimates, both the recomputations and the bootstrapping, led to improved performance, as indicated by the summary measures at the bottom of Table 6. The improvements were not dramatic, presumably because performance in these groups was reasonably accurate to start with. For the Partial Algorithm condition, there was little to distinguish between the two types of enhancement; for the Full Algorithm condition, arithmetic corrections of individual subjects' work produced more accurate estimates than did bootstrapping for 11 of the 16 questions.

Confidence in Accuracy of Estimation

In all estimation conditions and for each question, subjects were asked to give a rating on a 7-point scale indicating how confident they were in the accuracy of their answers (1 = "very unconfident"). Confidence ratings obtained in such a way are, of

course, fairly imprecise expressions of belief in the accuracy of estimation. They do, however, serve as a rough indication of the extent to which users believe in the accuracy of their results.

On average, subjects were generally unconfident in the accuracy of their estimates altogether. The mean confidence rating across all the data was only 2.97, perhaps reflecting the general difficulty of the task. The mean ratings were 2.84, 2.78, 3.01, 3.05, and 3.18 for the Unaided, List, List & Estimate, Partial Algorithm and Full Algorithm conditions respectively. An overall F-test on these results was significant, $\underline{F}$ $(4, 75) = 3.01$, $p < .05$. Individual t-tests indicated significant differences between the Full Algorithm condition and both the Unaided and the List conditions, $t(30) = 3.06$ and 3.12, $p < .01$. This indicates that providing greater degrees of problem structuring can result in a greater degree of confidence in the accuracy of estimates.

The increase in confidence with additional task structuring could have been due either to a general feeling that structure is bound to lead to more accurate answers or to a true sensitivity to actual improvements in the quality of estimates. Were the latter the case, one would expect to see moderately sized correlation coefficients within each of the estimation conditions between error ratios for each question and mean confidence ratings. Those correlations, however, tended to be positive but quite small. On average across groups, slightly less than 12% of the variance in

accuracy achieved in estimating the 16 problems was explained by the confidence judgments.

Discussion

The task used in the present study was quite difficult for our subjects (mostly college students). The answers provided by unaided subjects, when averaged over some 40 individuals, were in error by a factor of more than 10 for 7 of the 16 questions. Moreover, individual responses to a single question were enormously variable. For the unaided group, the largest response to a given question was, on average, 87 million times larger than the smallest response to the same question.

This poor performance was greatly improved by the structured aids we provided. For both accuracy and consistency, the results showed that, in general, the more structured the aid, the greater the improvement. With complete algorithmic decomposition (the most structured aid), the answers when averaged over 40 some individuals were always within a factor of 10 of the correct answer. This increased accuracy was accompanied by increased consistency; for the Full Algorithm group, the ratio of largest to smallest answer was, on average, 389,000. Readers who find even this range unduly large may take comfort from the interquartile range for the Full Algorithm group. On average, it was a respectable factor of 20, as compared with a factor of 56 for the unaided group.

Though the Partial Algorithm and Full Algorithm conditions produced more accurate and consistent estimates, they also involved

the use of problem representations constructed for the convenience of the study. The performance of those two approaches, therefore, is somewhat conditional on the representations we happened to choose. Since every estimation problem can have several representations, one has to be cautious in making general claims for the quality of these two approaches. What the results do suggest, however, is that people can perform estimation in this manner. More extensive training in the use of algorithms might improve performance; one might even hope for generalization of that training. For example, it is possible that an estimation training program based on algorithmic decomposition may generalize to subsequent performance on estimation problems when no problem representation is provided, similar to the List condition in the present study. Training in estimation using algorithmic formats may help stimulate a general mental model of estimation that could be drawn on in situations where no problem representation is otherwise available. Such models could be elicited and incorporated into decision support systems as an aid to users when the system is unable to provide the exact information called for.

An improvement in accuracy over direct estimation was also achieved by an approach in which subjects listed things they thought were important in making their estimates. A singular attribution for these results is difficult to make. Perhaps subjects simply spent longer on the List task and took more care in making their estimates than did those who were asked to estimate

the quantities directly. The act of listing knowledge they felt relevant to making an estimate may have facilitated retrieval from memory of facts they knew about the quantity in question. Lastly, the procedure may have provided a format for organizing what they knew about the topic in a more efficient way than holding everything in memory. All of these factors may have contributed some part to the effectiveness of the approach; further research is needed to explicate the effective mechanisms.

The failure of the List & Estimate approach to result in better quality estimates came as a surprise. In principle, asking individuals to develop their own structuring approach should have provided them the benefits of a decompositional strategy in the framework of a problem representation that bore some compatibility with their own. Secondarily, we anticipated that estimating values in one's own list would focus people more directly on the magnitude of the target number they were being asked to estimate. In retrospect, it appears likely that the effect of both creating a listing of relevant quantities and providing estimates for each item was to distract people's thinking from a coherent representation of the problem they were trying to solve. They may have become so involved in producing quantities and estimates that they were incapable of seeing the relationship between that task and the task of estimating the target quantity. The listing-only condition (List) was much less focused and gave people greater opportunity to think more generally about the problem. Without the

additional estimation tasks, they were less mentally burdened and perhaps more able to think wholistically about what they were doing.

It is possible that this condition prompted subjects to think in broad, associative terms about the quantity in question, somewhat akin to current theories of mental process that emphasize general activation of semantic memory in the performance of cognitive tasks (e.g., Collins & Loftus, 1975; Rumelhart & McClelland, 1986). No attempt was made in this study to control the specifics of the strategy that subjects used. Directing subjects to use a more restrictive line of thinking in producing their lists for this condition may have yielded an additional increment in estimation quality over that obtained here, especially if the instructions prompted them to generate at least partial algorithms.

These are, of course, hypotheses that are testable in future experiments. The relative effects on estimation accuracy of information processing load could be examined by calling upon subjects to provide degrees of structuring for estimation problems outside of the range they were required to do here. For example, a greater burden could be placed on their information processing capacity by asking them not only to make intermediate estimates, but to form them into an algorithm as well.

These results are suggestive of some potentially important design considerations in structuring knowledge retrieval. A

central dimension along which estimation approaches were constructed for this study was the degree of problem structuring provided to subjects, ranging from no structuring at all to highly structured algorithms. Each of those approaches varied as well in the information processing demands placed on users' mental resources. When specific problem representations were provided, the potential decrement to estimation quality due to the mentally taxing nature of the task may have been offset by the advantage of working within a format that handled significant details of the task such as organization and information integration. Freed from those mental chores, subjects were perhaps able to give greater attention to the task of generating estimates. Where no problem representation was provided and only minimal direction given to construct one, subjects were relatively free to concentrate predominantly on the task of estimating the target quantity. Introducing additional mental work by requiring individuals to provide both a specific problem structure and additional estimates of subquantities seemed to have the effect of reducing overall performance. Designers of approaches to facilitate knowledge retrieval may need to give careful attention to the psychological demands those approaches place upon the individuals for whom they are designed.

References

Armstrong, J. S., Denniston, W. B. Jr., & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. Organizational Behavior and Human Performance, 14, 257-263.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82, 407-428.

Dawes, R. M., & Corrigan, B. (1974). Linear Models in decision making. Psychological Bulletin, 81, 95-106.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 7, 86-106.

Falk, R. (1984). Multiplicative analogues of some statistics. American Mathematical Monthly, 91, 198-202.

Goodman, S. E., & Hedetniemi, S. T. (1977). Introduction to the design and analysis of algorithms. New York: McGraw Hill, Inc.

Hammond, K. (1981). Principles of organization in intuitive and analytical cognition (Report No. 231). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Hammond, K., Hamm, R. M., Grassia, J., & Pearson, T. (1983). Direct comparison of intuitive, quasi-rational and analytical cognition (Report No. 248). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Horabin, I., & Lewis, B. (1978). The instructional design library [Vol. 2: Algorithms]. Englewood Cliffs, NJ: Educational Technology Pub. Co.

Landa, L. N. (1976). Instructional regulation and control: Cybernetics, algorithmization, and heuristics in education. Englewood Cliffs, NJ: Educational Technology Pub. Co.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance, 20, 159-183.

Lyness, K. S., & Cornelius, E. T. (1982). A comparison of holistic and decomposed judgment strategies in a performance rating simulation. Organizational Behavior and Human Performance, 29, 21-38.

Peters, J. T., Hammond, K. R., & Summers, D. A. (1974). A note on intuitive vs. analytic thinking. Organizational Behavior and Human Performance, 12, 125-131.

Raiffa, H. (1968). Decision analysis. Reading, MA: Addison-Wesley.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel distributed processing. Cambridge, MA: The MIT Press.

Singer, M. (1971). The vitality of mythical numbers. The Public Interest, Spring, 3-9.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 4, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Wright, P., & Reid, F. (1973). Written information: Some alternatives to prose for expressing the outcome of complex contingencies. Journal of Applied Psychology, 57, 160-166.

Author Notes

This research was supported by the Army Research Institute under contract MDA 903-83-C-0198.

Requests for reprints should be sent to Donald MacGregor, Decision Research, 1201 Oak Street, Eugene, OR 97401.

Footnotes

[1] We recognize that these "correct answers" may contain errors. However, we believe that these errors are slight when compared to the size of errors made by the subjects.

[2] A common index of consistency is the standard deviation of a distribution. That metric, however, is an inappropriate index of variability for distributions of logs (see Falk, 1984). The expression of variability used here was chosen because it is easily interpretable and is a permissible index given the data on which it is computed.

Table 1

Estimation Questions and Correct Answers

Question	Correct Answer
Number of practicing physicians in Lane County.	350
Marriage licenses issued in Lane County last year.	2,486
Alcoholics, 20 years old or older, in Oregon last year.	34,000
Ph.D.'s granted in all fields in the U. S. last year.	34,086
Forested square miles in Oregon.	47,506
People employed by colleges and universities in the U. S. last year.	1,935,000
Tons of fish caught by U.S. commercial fishermen last year	2,421,000
People in the U.S employed by hospitals last year.	3,568,000
Gallons of water used daily by all homes in Eugene, OR.	6,135,000
Total attendance at all regular season major-league baseball games last year.	31,318,000
Total welfare payments to Oregon families with dependent children (AFDC) last year.	$117 million

(Table continues)

Table 1, continued

Question	Correct Answer
Value of all new imported passenger cars sold in the U. S. last year.	$12 billion
Dollars spent in the U.S. last year on alcoholic beverages (beer, wine & liquor) for personal consumption.	$24.7 billion
Pieces of mail handled by the U.S. Postal Service last year.	89.8 billion
Gallons of fuel consumed by U.S. motor vehicles last year.	109 billion
Cigarettes consumed in the U.S. last year.	604.1 billion

Table 2. Abbreviated Descriptions of Algorithms

PHYSICIANS
County population
Number of physician visits per person per year
Hours per week the average physician works
Proportion physician work hours seeing patients
Weeks per year physician works
Length of average doctor visit

ALCOHOLICS
Number of women age 20 and over in Oregon
Proportion of Oregon women 20 and over who are alcoholic
Number of men age 20 and over in Oregon
Proportion of Oregon men 20 and over who are alcoholic

FORESTED MILES
Distance between north & south state borders
Distance between west & east state borders
Proportion of Oregon that is forested

FISH
Number of U.S. ports with commercial fishing boats
Average number of fishing boats per port
Tons of fish caught per trip per boat
Number of trips per year per boat

HOSPITAL EMPLOYEES
Average number of hospitals per state
Number of states
Average number of employees per hospital

MARRIAGES
County population
Proportion of population of marriageable age
Proportion of marriageable-age population marrying each year
Number of people in a marriage

PH.D.'S
Average number of universities granting Ph.D.'s per state
Number of states
Number of Ph.D.'s granted by the average university

UNIVERSITY EMPLOYEES
Average number of universities per state
Number of states
Average number of employees per university

WATER
Eugene population
Gallons of water used per day by average person:
- in the bathroom
- in the kitchen
- cleaning inside the house
- outside the house

BASEBALL
Number of major-league baseball teams
Number of regular-season games per team
Number of teams in a single game
Average attendance per baseball game

(Table continues)

Table 2, continued

Table 2, continued

WELFARE Oregon population Proportion of population receiving welfare payments Average monthly dollar payment per welfare recipient Months in a year	**POST OFFICE** Average number of post offices per state Number of states Pieces of mail per day handled by average post office Days in a year
IMPORTED CARS Population of the U.S. Number of passenger cars per person in the U.S. Proportion of passenger cars purchased new each year Proportion of new passenger cars that are imported Average dollar value of per imported passenger car	**GASOLINE** Number of cars in U.S. Average number of miles driven per car per year Average number of miles per gallon per car Number of buses in U.S. Average number of miles driven per bus per year Average number of miles per gallon per bus Number of trucks in U.S. Average number of miles driven per truck per year Average number of miles per gallon per truck
ALCOHOL DOLLARS Alcohol consumed per average person per month: Number of cans of beer Number of bottles of wine Number of bottles of liquor Average cost for alcoholic beverages: Per can of beer Per bottle of wine Per bottle of liquor Months in a year Population of the U.S.	**CIGARETTES** Population of the U.S. Proportion of population that smokes Number of cigarettes consumed per day by average smoker Days in a year

Table 3

Summary of results: Geometric mean estimates.

Question	Unaided	List	List & Estimate	Partial Algorithm	Full Algorithm	Correct Answer
Physicians	456	976	691	238	421	350
Marriages	2,348	2,920	2,553	4,897	7,648	2,486
Alcoholics	17,583	26,556	34,473	30,683	120,319	34,000
Ph.D.'s	20,903	22,560	32,937	12,706	27,596	34,086
Forested Miles	31,945	38,637	6,267	31,405	45,097	47,506
Univ. Employees	1,053,483	388,088	411,476	233,776	246,589	1,935,000
Fish	660,187	2,082,276	180,225	882,267	7,289,706	2,421,000
Hosp. Employees	1,123,043	1,373,184	753,169	874,883	935,398	3,568,000

(Table continues)

Table 3, Continued

Question	Unaided	List	List & Estimate	Partial Algorithm	Full Algorithm	Correct Answer
Water	61,070	551,756	234,097	2,516,518	2,728,545	6,135,000
Baseball	4,003,923	11,150,029	3,376,414	10,322,859	4,183,163	31,318,000
Welfare	6,556,648	3,063,380	6,066,844	13,278,224	605,290,000	117,000,000
Imported Cars	669,160,200	177,540,770	84,277,782	239,720,000	3,012,200,000	12,000,000,000
Alcohol Dollars	223,833,460	123,674,570	330,850,000	1,516,700,000	7,433,100,000	24,700,000,000
Post Office	1,004,650,700	2,529,108,700	5,027,300,000	12,023,000,000	10,159,000,000	89,800,000,000
Gasoline	268,726,690	3,713,726,400	22,354,000,000	22,856,000,000	10,951,000,000	109,000,000,000
Cigarettes	1,536,056,200	7,068,404,600	1,011,300,000	401,240,000,000	73,173,000,000	604,100,000,000

Table 4

Summary of Results: Error Ratios

Question	Unaided	List	List & Estimate	Partial Algorithm	Full Algorithm
Physicians	1.30 (+)[a]	2.79 (+)	1.97 (+)	1.47 (-)	1.20 (+)
Marriages	1.06 (-)	1.17 (+)	1.03 (+)	1.97 (+)	3.08 (+)
Alcoholics	1.93 (-)	1.28 (-)	1.01 (+)	1.11 (-)	3.54 (+)
Ph.D.'s	1.63 (-)	1.51 (-)	1.03 (-)	2.86 (-)	1.24 (-)
Forested Miles	1.49 (-)	1.23 (-)	7.58 (-)	1.51 (-)	1.05 (-)
Univ. Employees	1.84 (-)	4.99 (-)	4.70 (-)	8.28 (-)	7.85 (-)
Fish	3.67 (-)	1.16 (-)	13.43 (-)	2.74 (-)	3.01 (-)
Hosp. Employees	3.18 (-)	2.60 (-)	4.74 (-)	4.08 (-)	3.81 (-)
Water	100.46 (-)	11.12 (-)	26.21 (-)	2.44 (-)	2.25 (-)
Baseball	7.82 (-)	2.81 (-)	9.28 (-)	3.03 (-)	7.49 (-)
Welfare	17.84 (-)	38.19 (-)	19.29 (-)	8.81 (-)	5.17 (+)
Imported Cars	17.93 (-)	67.59 (-)	142.39 (-)	50.06 (-)	3.98 (-)
Alcohol Dollars	110.35 (-)	199.68 (-)	74.66 (-)	16.29 (-)	3.32 (-)
Post Office	89.38 (-)	35.49 (-)	17.86 (-)	7.47 (-)	8.84 (-)
Gasoline	405.66 (-)	29.36 (-)	4.88 (-)	4.77 (-)	9.95 (-)
Cigarettes	393.29 (-)	85.47 (-)	597.35 (-)	1.51 (-)	8.26 (-)
Median	5.75	3.90	8.43	2.89	3.68
Mean	72.41	30.40	59.08	7.39	4.63

[a] (+) indicates overestimation; (-) indicates underestimation.

Table 5

Ranges and Interquartile Ranges of Log Estimates.

Question	Overall Range					Interquartile Range				
	Unaid	List	L&Est	PAlg	FAlg	Unaid	List	L&Est	PAlg	FAlg
Physicians	4.00	3.05	3.08	2.40	5.69	.64	1.20	.92	.89	1.14
Marriages	4.01	3.78	3.02	4.00	5.25	1.17	.60	.93	1.01	1.07
Alcoholics	4.87	3.70	3.52	2.48	5.49	1.28	1.00	1.05	1.00	1.19
Ph.D.'s	7.98	6.88	6.30	3.12	2.90	1.57	1.82	1.10	1.19	.70
Forested Miles	7.90	7.85	3.99	5.70	4.54	2.30	1.41	2.05	1.46	.80
Univ. Employees	10.26	4.30	5.51	5.00	7.70	1.18	1.66	1.24	1.01	1.36
Fish	13.00	9.92	9.82	10.31	7.48	2.15	2.93	3.30	3.12	2.01
Hosp. Employees	7.00	5.00	5.40	5.27	3.41	1.10	.90	1.32	1.15	1.40
Water	7.08	8.95	6.30	3.41	4.86	1.75	1.48	1.22	.74	.62
Baseball	8.52	5.40	4.75	6.00	3.74	1.43	1.18	1.49	3.00	1.24
Welfare	7.07	5.30	5.60	6.01	6.13	1.75	1.50	1.90	2.65	.78
Imported Cars	9.97	8.26	7.76	7.25	6.88	2.06	2.93	3.05	1.92	2.11
Alcohol Dollars	12.00	4.78	8.30	6.51	7.08	2.60	2.00	2.86	1.22	1.59
Post Office	11.70	7.40	11.60	10.30	8.48	2.00	1.73	2.30	2.26	2.47
Gasoline	7.78	6.00	8.30	8.93	8.56	2.76	2.90	1.60	2.63	2.76
Cigarettes	19.00	9.64	9.09	7.85	5.23	2.52	1.68	2.80	1.94	1.47
Median	7.94	5.70	6.31	5.85	5.59	1.75	1.58	1.55	[illegible]	[illegible]

Table 6

Error Ratios for Subjects' Responses and Recomputed Responses

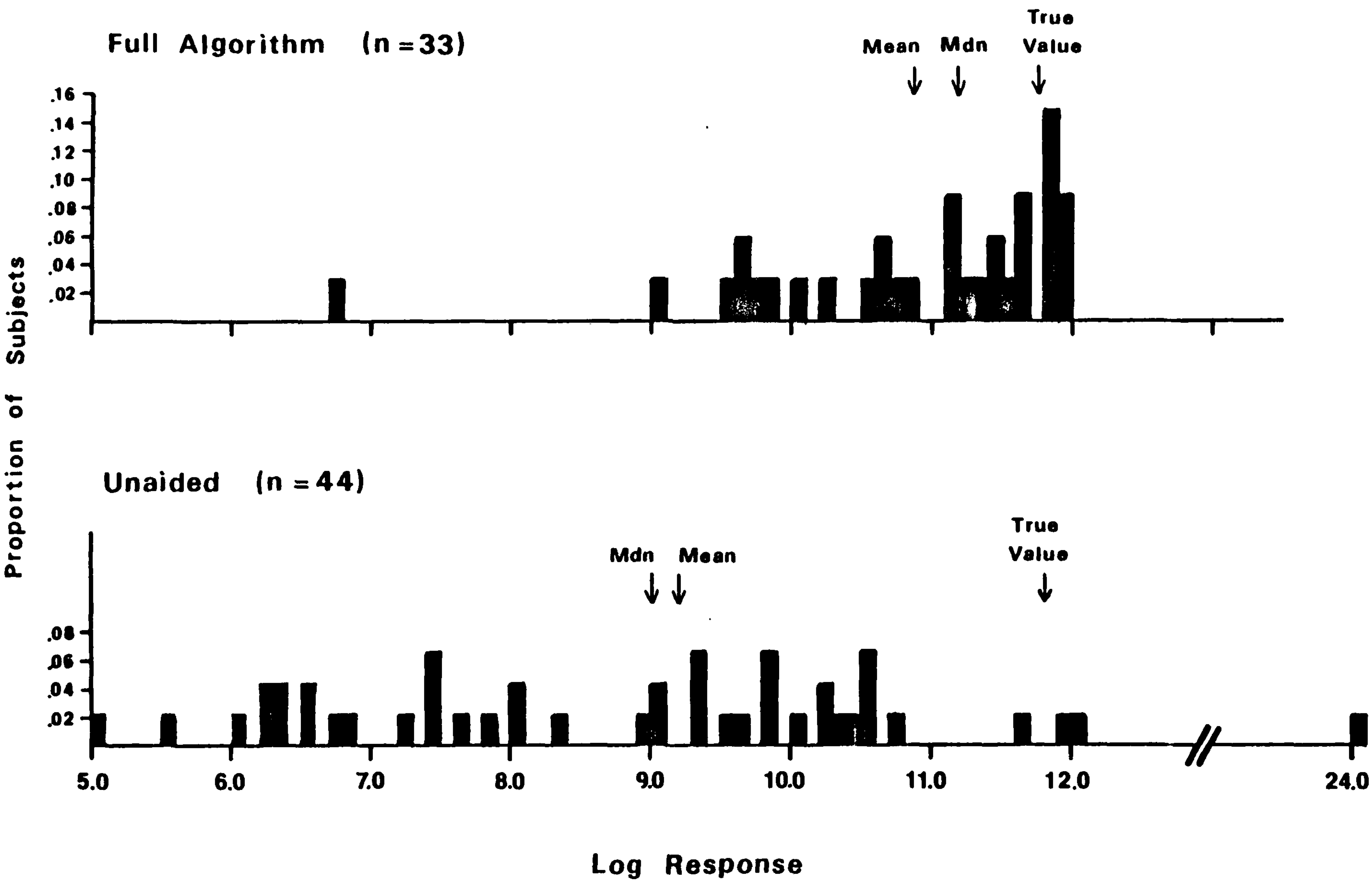
	Partial Algorithm			Full Algorithm		
Question	Estimate	Computed	Boot[a]	Estimate	Computed	Boot[a]
Physicians	1.47(-)[b]	2.89(-)	2.41(-)	1.20(+)	2.13(-)	2.20(-)
Marriages	1.97(+)	3.26(+)	3.52(+)	3.08(+)	2.03(+)	4.02(+)
Alcoholics	1.11(-)	2.61(+)	4.06(+)	3.54(+)	4.15(+)	7.35(+)
Ph.D.'s	2.68(-)	1.17(-)	1.36(-)	1.24(-)	1.06(-)	1.82(-)
Forested Miles	1.51(-)	1.26(+)	1.09(+)	1.05(-)	1.75(+)	1.60(+)
Univ. Employee	8.28(-)	3.92(-)	7.74(-)	7.85(-)	4.17(-)	4.85(-)
Fish	2.74(-)	20.44(+)	7.23(+)	3.01(+)	3.97(+)	4.65(+)
Hosp. Employees	4.08(-)	2.18(-)	2.85(-)	3.81(-)	2.28(-)	2.71(-)
Water	2.44(-)	2.06(-)	1.64(-)	2.25(-)	1.26(-)	1.53(-)
Baseball	3.03(-)	2.24(-)	1.42(-)	7.49(-)	4.27(-)	5.52(-)
Welfare	8.81(-)	2.11(+)	2.87(+)	5.17(+)	5.55(+)	4.10(+)
Imported Cars	50.06(-)	1.44(+)	3.85(+)	3.98(-)	3.73(+)	6.72(+)
Alcohol Dollars	16.29(-)	2.76(-)	2.81(-)	3.32(-)	2.74(-)	1.27(-)
Post Office	7.47(-)	1.82(+)	1.09(-)	8.84(-)	1.05(-)	3.28(-)
Gasoline	4.77(-)	1.47(-)	1.08(-)	9.95(-)	3.21(-)	2.28(-)
Cigarettes	1.51(-)	1.23(-)	1.12(-)	8.26(-)	2.00(-)	1.19(-)
Median	2.89	2.15	2.61	3.68	2.51	3.00
Mean	7.39	3.30	2.88	4.63	2.83	3.44

[a]Bootstrap values were computed using medians of subjects' component estimates.

[b](+) indicates overestimation; (-) indicates underestimation.

Figure Caption

Figure 1. Response distributions for two of the groups (Full Algorithm and Unaided) on the Cigarette question.


Full Algorithm (n = 33)
Mean
Mdn
True Value
.16
.14
.12
.10
.08
.06
.04
.02
Proportion of Subjects
Unaided (n = 44)
Mdn
Mean
True Value
.08
.06
.04
.02
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
24.0
Log Response